

# Willingness Optimization for Social Group Activity

Hong-Han Shuai  
National Taiwan Univ.

{hshuai,dnyang}@iis.sinica.edu.tw

De-Nian Yang  
Academia Sinica

Philip S. Yu  
Univ. of Illinois at Chicago

psyu@cs.uic.edu

Ming-Syan Chen  
National Taiwan Univ.

mschen@cc.ee.ntu.edu.tw

## ABSTRACT

Studies show that a person is willing to join a social group activity if the activity is interesting, and if some close friends also join the activity as companions. The literature has demonstrated that the interests of a person and the social tightness among friends can be effectively derived and mined from social networking websites. However, even with the above two kinds of information widely available, social group activities still need to be coordinated manually, and the process is tedious and time-consuming for users, especially for a large social group activity, due to complications of social connectivity and the diversity of possible interests among friends. To address the above important need, this paper proposes to automatically select and recommend potential attendees of a social group activity, which could be very useful for social networking websites as a value-added service. We first formulate a new problem, named Willingness mAximization for Social grOup (WASO). This paper points out that the solution obtained by a greedy algorithm is likely to be trapped in a local optimal solution. Thus, we design a new randomized algorithm to effectively and efficiently solve the problem. Given the available computational budgets, the proposed algorithm is able to optimally allocate the resources and find a solution with an approximation ratio. We implement the proposed algorithm in Facebook, and the user study demonstrates that social groups obtained by the proposed algorithm significantly outperform the solutions manually configured by users.

## 1. INTRODUCTION

Studies show that two important criteria are usually involved in the decision of a person joining a group activity [7, 12] at her available time. First, the person is interested in the intrinsic properties of the activity, which may be in line with her favorite hobby or exercise. Second, other people who are important to the person, such as her close friends, will join the activity as companions<sup>1</sup>. For example, if a person who appreciates abstract art has complimentary tickets for a modern art exhibition at MoMA, she would

<sup>1</sup>There are other criteria that are also important, e.g., activity time, and activity location. However, to consider the above factors, a promising way is to preprocess and filter out the people who are not available, live too far, etc.

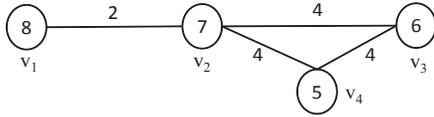
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/>. Obtain permission prior to any use beyond those covered by the license. Contact copyright holder by emailing [info@vlldb.org](mailto:info@vlldb.org). Articles from this volume were invited to present their results at the 40th International Conference on Very Large Data Bases, September 1st - 5th 2014, Hangzhou, China. *Proceedings of the VLDB Endowment*, Vol. 7, No. 4. Copyright 2013 VLDB Endowment 2150-8097/13/12.

probably want to invite her friends and friends of friends with this shared interest. Nowadays, many people are accustomed to sharing information with their friends on social networking websites, like Facebook, Myspace, Meetup, MyYearbook, and LikeALittle, and a recent line of studies [4, 15] has introduced effective algorithms to quantify the interests of a person according to the interest attributes in her personal profile and the contextual information in her interaction with friends. Moreover, social connectivity models have been widely studied [2] for evaluating the tightness between two friends in the above websites. Nonetheless, even with the above knowledge available, to date there has been neither published work nor a real system explores how to leverage the above two crucial factors for *automatic planning and recommending of a group activity*, which is potentially very useful for social networking websites as a value-added service<sup>2</sup>. At present, many social networking websites only act as a platform for information sharing and exchange in activity planning. The attendees of a group activity still need to be selected manually, and such manual coordination is usually tedious and time-consuming, especially for a large social activity, given the complicated social link structure and the diverse interests.

To solve this problem, this paper makes an initial attempt to incorporate the interests of people and their social tightness as two key factors to find a group of attendees for automatic planning and recommendation. It is desirable to choose more attendees who like and enjoy the activity and to invite more friends with the shared interest in the activity as companions. In fact, Psychology [7, 12] and recent study in social networks [21, 22] have modeled the *willingness* to attend an activity or a social event as the sum of the interest of each attendee on the activity and the social tightness between friends that are possible to join it. It is envisaged that the selected attendees are more inclined to join the activity if the willingness of the group increases.

With this objective in mind, we formulate a new fundamental optimization problem, named *Willingness mAximization for Social grOup (WASO)*. The problem is given a social graph  $G$ , where each node represents a candidate person and is associated with an interest score of the person for the activity, and each edge has a social tightness score to indicate the mutual familiarity between the two persons. Let  $k$  denote the number of expected attendees. Given the user-specified  $k$ , the goal of automatic activity planning is to maximize the willingness of the selected group  $F$ , while the induced graph on  $F$  is a connected subgraph for each attendee to become

<sup>2</sup>The privacy of a person in automatic activity planning can follow the current privacy setting policy in social networking websites when the person subscribes the service. The details of privacy setting are beyond the scope of this paper.



**Figure 1: Counterexample of greedy algorithm**

acquainted with another attendee according to a social path<sup>3</sup>. For the activities without an *a priori* fixed size, it is reasonable for a user to specify a proper range for the group size, and our algorithm can find the solution for each  $k$  within the range and return the solutions for the user to decide the most suitable group size and the corresponding attendees.<sup>4</sup>

Naturally, to incrementally construct the group, a greedy algorithm sequentially chooses an attendee that leads to the largest increment in the willingness at each iteration. For example, Figure 1 presents an illustrative example with  $k = 3$ . Node  $v_1$  is first selected since its interest score is the maximum one among all nodes. Afterward, node  $v_2$  is then extracted. Finally,  $v_3$ , instead of  $v_4$ , is chosen because it generates the largest increment on willingness, i.e., 10, and leads to a group with a willingness of 27. Note the greedy algorithm, though simple, tends to be trapped in a local optimal solution, since it facilitates the selection of nodes only suitable at the corresponding iterations. In this simple example, the above algorithm is not able to find the optimal solution because it makes a greedy selection at each iteration and only chooses  $v_1$  as the start node, who enjoys the activity the most at the first iteration, but the optimal solution is  $\{v_2, v_3, v_4\}$  with the total willingness being 30.

Another approach is to examine the willingness of every possible combination of  $k$  attendees. However, this enumeration approach needs to evaluate  $C_k^n$  candidate groups, where  $n$  is the number of nodes in  $G$ . In current social networking websites, the number of candidate groups is still huge even when we focus on only the candidates located in the same area, e.g., about ten thousand users in British Virgin Islands<sup>5</sup>. When  $k = 50$ , the number of candidate groups is in the order of  $10^{135}$ . Thus, this approach is computationally intractable for a massive social network.

Indeed, we show that the problem is challenging and prove that it is NP-hard. As shown in Figure 1, the greedy approach improperly chooses  $v_1$  as the start node and explores only a single sequence of nodes in the solution space. To increase the search space, randomized algorithms have been proposed as a simple but effective strategy to solve the problems with large instances [16]. To avoid being trapped in a local optimal solution, a simple randomized algorithm for WASO is to randomly choose multiple start nodes. Each start node is considered as partial solution, and a node neighboring the partial solution is randomly chosen and added to the partial solution at each iteration afterward, until  $k$  nodes are included as a final solution. This randomized algorithm is more efficient than the greedy approach, because the computation of willingness is not involved during the selection of a node. For the problem with a large  $k$ , numerous candidate nodes neighboring the partial solution are necessary to be examined in the greedy approach to sum up the willingness, in order to find the one that generates the largest willingness. In contrast, the randomized algorithm simply chooses one neighboring node at random.

<sup>3</sup>For some group activities, it is not necessary to ensure that the solution group is a connected subgraph. Later in Section 2, we will show that WASO without a connectivity constraint can be easily solved by the proposed algorithm with simple modification.

<sup>4</sup>The parameter settings in WASO to fit varied scenarios in everyday life will be introduced in more details in Section 2.2.

<sup>5</sup><http://www.socialbakers.com/facebook-statistics/>.

With randomization, the aforementioned algorithm is able to effectively avoid being trapped in a local optimal solution. It suffers, however, two disadvantages. Firstly, a start node that has the potential to generate final solutions with high willingness is not invested with more computational budgets for randomization in the following iterations. Each start node in the randomized algorithm is expanded to only one final solution. Thus, a start node, which has the potential to grow and become the solution with high willingness, may fail to generate a final solution with high willingness because only one solution is randomly expanded from the start node. The second disadvantage is that the expansion of the partial solution does not differentiate the selection of the neighboring nodes. Each neighboring node is treated equally and chosen uniformly at random for each iteration. In contrast, a simple way to remedy this issue is to assign the probability to each neighboring node according to its interest score and social tightness of incident edges. However, this assignment is similar to the greedy algorithm in that it limits the scope to the local information corresponding to each node and is not expected to generate a solution with high willingness.

Keeping in mind the above observations in an effort to guide an efficient search of the solution space, we propose two randomized algorithms, called *CBAS* (Computational Budget Allocation for Start nodes) and *CBAS-ND* (Computation Budget Allocation for Start nodes with Neighbor Differentiation), to address the above two crucial factors in selecting start nodes and expanding the partial solutions, respectively. This paper exploits the notion of Optimal Computing Budget Allocation (OCBA) [3] in randomization, in order to optimally invest more computational budgets in the start nodes with the potential to generate the solutions with high willingness. *CBAS* first selects  $m$  start nodes and then randomly adds neighboring nodes to expand the partial solution stage-by-stage, until  $k$  nodes are included as a final solution. Each start node in *CBAS* is expanded to multiple final solutions. To properly invest the computational budgets, *CBAS* at each stage identifies the start nodes worth more computational budgets according to sampled results of the previous stages. Equipped with the allocation strategy of computational resources, *CBAS* is enhanced to *CBAS-ND* to adaptively assign the probability to each neighboring node during the expansion of the partial solution according to the cross entropy method. We prove that the allocation of computational budgets for start nodes and the assignment of the probability to each node are both optimal in *CBAS* and *CBAS-ND*, respectively. We further show that *CBAS* can achieve an approximation ratio, while *CBAS-ND* needs much smaller computational budgets than *CBAS* to acquire the same solution quality.

The contributions of this paper can be summarized as follows.

- We formulate a new optimization problem, namely WASO, to consider the topic interest of users and social tightness among friends for automatic planning of activities. We prove that WASO is NP-hard. To the best of the authors' knowledge, there is no real system or existing work in the literature that addresses the issue of automatic activity planning based on both topic interest and social relationship.
- We design Algorithm *CBAS* and *CBAS-ND* to find the solution to WASO with an approximation ratio. Experimental results demonstrate that the solution returned by *CBAS-ND* is very close to the optimal solution obtained by IBM CPLEX, which is widely regarded as the fastest general parallel optimizer, and *CBAS-ND* is faster than CPLEX.
- We implement *CBAS-ND* in Facebook and conduct a user study with 137 people. Currently, people are used to organizing an activity manually without being aware of the quality of

the organized group, because there is no automatic group recommendation service available for comparison. Compared with the manual solutions, we observe that the solutions obtained by *CBAS-ND* are 50.6% better. In addition, 98.5% of users conclude that the group recommended by *CBAS-ND* is better or acceptable. Therefore, this research result has the potential to be adopted in social networking websites as a value-added service.

The rest of this paper is summarized as follows. Section 2 formulates WASO and surveys the related works. Sections 3 and 4 explain *CBAS* and *CBAS-ND* and derive the approximation ratio. User study and experimental results are presented in Section 5, and we conclude this paper in Section 6.

## 2. PRELIMINARY

### 2.1 Problem Definition

Given a social network  $G = (V, E)$ , where each node  $v_i \in V$  and each edge  $e_{i,j} \in E$  are associated with an interest score  $\eta_i$  and a social tightness score  $\tau_{i,j}$  assigned according to the literature [2] and [4] respectively, this paper studies a new optimization problem WASO for finding a set  $F$  of vertices with size  $k$  to maximize the willingness  $W(F)$ , i.e.,

$$\max_F W(F) = \max_F \sum_{v_i \in F} (\eta_i + \sum_{v_j \in F: e_{i,j} \in E} \tau_{i,j}), \quad (1)$$

where  $F$  is a connected subgraph in  $G$  to encourage each attendee to be acquainted with another attendee according to a social path in  $F$ . Notice that the social tightness between  $v_i$  and  $v_j$  is not necessarily symmetric, i.e.,  $\tau_{i,j}$  can be different with  $\tau_{j,i}$ . Therefore, the willingness in Eq. (1) considers both  $\tau_{i,j}$  and  $\tau_{j,i}$ <sup>6</sup>. As demonstrated in the previous works in Psychology and social networks [21, 22] that jointly consider the social and interest domains, the willingness of a group is represented as the sum of the topic interest of nodes and social tightness between them<sup>7</sup>.

Notice that the network with  $\eta$  as 0 or  $\tau$  as 0 is a special case of WASO. Previous works [7, 12] demonstrated that both social tightness and topic interest are intrinsic criteria involved in the decision of a person to join a group activity, which is in line with the results of our user study presented in Section 5. WASO is challenging due to the tradeoff in interest and social tightness, while the constraint that assures that the  $F$  is connected also complicates this problem because it is no longer able to arbitrarily choose any nodes from  $G$ . Indeed, the following theorem shows that WASO is NP-hard.

**THEOREM 1.** *WASO is NP-hard.*

**PROOF.** We prove that WASO is NP-hard with the reduction from Dense  $k$ -Subgraph (DkS) [8]. Given a graph  $G_D = (V_D, E_D)$ , DkS finds a subgraph with  $k$  nodes  $F_D$  to maximize the density of the subgraph. In other words, the purpose of DkS is to maximize the number of edges  $E(F_D)$  in the subgraph induced by the selected nodes. For each instance of DkS, we construct an instance for WASO by letting  $G = G_D$ , where  $\eta_i$  of each node  $v_i \in V$  is set as 0, and  $\tau_{i,j}$  of each edge  $e_{i,j} \in E$  is assigned as 1. We first prove the sufficient condition. For each instance of DkS with solution node set  $F_D$ , we let  $F = F_D$ . If the number of edges  $E(F_D)$  in the subgraph of DkS is  $\epsilon$ , the willingness of WASO  $W(F)$  is also  $\epsilon$  because  $F = F_D$ . We then prove the necessary condition.

<sup>6</sup>The illustrating examples are symmetric for simplicity.

<sup>7</sup>Different weights  $\lambda$  and  $(1-\lambda)$  can be assigned to the interest scores and social tightness such that  $W(F) = \sum_{v_i \in F} (\lambda_i \eta_i + (1 - \lambda_i) \sum_{v_j \in F: e_{i,j} \in E} \tau_{i,j})$ .  $\lambda_i$  can be set directly by a user or according to the existing model [22]. The impacts of different  $\lambda$  will be studied later in Section 5.

For each instance of WASO with  $F$ , we select the same nodes for  $F_D$ , and the number of edges  $E(F_D)$  must be the same as  $W(F)$ . The theorem follows.  $\square$

### 2.2 Scenarios

In the following, we present the parameter settings of WASO to fit the need of different scenarios.

*Couple and Foe:* For any two people required to be selected together, such as a couple, the two corresponding nodes  $v_i$  and  $v_j$  in  $G$  are merged into one node  $v_a$  with the interest score  $\eta_a = \eta_i + \eta_j$  and social tightness score  $\tau_{a,b} = \tau_{i,b} + \tau_{j,b}$  for each neighboring node  $v_b$  of  $v_i$  or  $v_j$ . Similarly, more people can be merged as well, but the group size  $k$  is required to be adjusted accordingly. On the other hand, if  $v_i$  is a foe of  $v_j$ , their social tightness score  $\tau_{j,i}$  is assigned a large negative value, such that any group consisting of the two nodes leads to a negative willingness and thereby will not be selected. The relationship of foes can be discovered by blacklists and learnt from historical records. Similarly,  $\eta_i$  is allowed to be assigned a negative value.

*Invitation:* A piano player plans to hold a small concert. In this case, the player might prefer inviting people that are very good friends with him/her, but it is not necessary for them to be pair-wise acquainted. For this scenario, the activity candidates are the neighboring nodes of  $v_i$ , which is denoted as  $N(v_i)$ , where  $v_i$  represents the inviter (piano player), and we set  $\lambda_j$  as 1 for every  $j \in N(v_i)$  since the social tightness among the friends may not be important in this scenario.

*Exhibition and House-Warming Party:* The British Museum plans to hold an exhibition of Van Gogh and would like to send e-mails to potential visitors. In this scenario, the topic interest is expected to play a crucial role, and  $\lambda_i$  is suitable to set as 1 for all  $i \in V$ . On the other hand, for social activities such as a house-warming party,  $\lambda_i$  is 0 for all  $i \in V$ , and only social tightness is considered.

*Separate Groups:* The government plans to organize a camping trip on Big Bear Lake to promote environmental protection. In this case, the group does not need to be connected, and a simple way is to add a virtual node  $\bar{v}$  to  $V$  with the interest score of  $\bar{v}$  as  $\eta_{\bar{v}} = \epsilon + \sum_{v_i \in V} (\eta_i + \sum_{v_j \in V: e_{i,j} \in E} \tau_{i,j})$ , where  $\epsilon$  is any positive real number. In addition,  $\bar{v}$  is connected to every other node  $v_j \in V$  with the social tightness score  $\tau_{\bar{v},j} = 0$ , and the set of new edges incident to  $\bar{v}$  is denoted as  $E_{\bar{v}}$ . It is necessary to choose  $\bar{v}$  so that  $\bar{v}$  will connect to multiple disconnected subgraphs to support the above group activities. In this case,  $k + 1$  nodes need to be included in the final solution.

Now let WASO-dis denote the counterpart of WASO without the connectivity constraint. Indeed, WASO-dis is simpler than WASO because the constraint is not incorporated. In the following, we prove that WASO can be reduced from WASO-dis. In other words, any algorithm for WASO can also solve WASO-dis.

**THEOREM 2.**  *$F_d^*$  is the  $k$ -node optimal solution of any WASO-dis instance  $G_d$  if and only if  $F^*$  is the  $k + 1$ -node optimal solution of WASO instance  $G$ , where  $F^* = F_d^* \cup \bar{v}$ .*

**PROOF.** We first prove the sufficient condition. Since the optimal solution of WASO must include  $\bar{v}$ , if  $F_d^* \cup \bar{v}$  is not optimal to WASO, there exists a better solution  $\bar{F}$  with  $W(\bar{F}) \geq W(F_d^* \cup \bar{v})$ , which implies that  $W(\bar{F} - \bar{v}) > W(F_d^*)$ . Because  $\bar{F} - \bar{v}$  can act as a feasible solution to WASO-dis,  $W(\bar{F} - \bar{v}) > W(F_d^*)$  contradicts that  $F_d^*$  is optimal to WASO-dis. Therefore,  $F^* = F_d^* \cup \bar{v}$  is the optimal solution to WASO.

We then prove the necessary condition. Since the optimal solution of WASO must include  $\bar{v}$ , if  $F^* - \bar{v}$  is not optimal to WASO-dis, there exists a better solution  $\bar{F}_d$  with  $W(\bar{F}_d) > W(F^* - \bar{v})$ ,

implying that  $W(\overline{F}_d + \overline{v}) > W(F^*)$ , contradicting that  $F^*$  is optimal to WASO because  $\overline{F}_d + \overline{v}$  is also a feasible solution to WASO. The theorem follows.  $\square$

### 2.3 Related Works

Given the growing importance of varied social networking applications, there has been a recent push on the study of user interest scores and social tightness scores from real social networking data. It has been demonstrated that unknown user interest attributes can be effectively inferred from a social network according to the revealed attributes of the friends [15]. On the other hand, Wilson *et al.* [20] derived a new model to quantify the social tightness between any two friends in Facebook. The number of wall postings is also demonstrated to be an effective indicator for social tightness [9]. Thus, the above studies provide a sound foundation to quantify the user interest and social tightness scores in social networks. Moreover, Yang [21] and Lee [22] sum up the two factors as willingness for marketing and recommendation. Nevertheless, the above factors crucial in social networks have not been leveraged for automatic activity planning explored in this paper.

Expert team formation in social networks has attracted extensive research interests. The problem of constructing an expert team is to find a set of people owning the specified skills, while the communications cost among the chosen friends is minimized to ensure the rapport among the team members for an efficient operation. Two communications costs, diameter and minimum spanning tree, were evaluated. Several extended models have been studied. For example, each skill  $i$  needs to contain at least  $k_i$  people in order to form a strong team [14], while all-pair shortest paths are incorporated to describe the communications costs more precisely [13]. Moreover, a skill leader is selected for each skill with the goal to minimize the social distance from the skill members to each skill leader.

In addition to expert team formation, community detection as well as graph clustering and graph partitioning has been explored to find groups of nodes mostly based on the graph structure [1]. The quality of an obtained community is usually measured according to the structure inside the community, together with the connectivity within the community and between the rest of the nodes in the graph, such as the density of local edges, deviance from a random null model, and conductance [10]. Sozio *et al.* [19], for example, detected community by minimizing the total degree of a community with specified nodes. However, the objective function of WASO is different from community detection. Each node and each edge in WASO are associated with an interest score and social tightness score in the problem studied in this paper, in order to maximize the willingness of the attendees with a specified group size, which can be very useful for social networking websites as a value-added service.

## 3. ALGORITHM DESIGN FOR WASO

To solve WASO, a greedy approach incrementally constructs the solution by sequentially choosing an attendee that leads to the largest increment in the willingness at each iteration. However, while this approach is simple, it tends to be trapped in a local optimal solution. The search space of the greedy algorithm is limited because only a single sequence of nodes is explored. To address the above issues, this paper first proposes a randomized algorithm *CBAS* to randomly choose  $m$  start nodes. Each start node acts as a seed to be expanded to multiple final solutions. At each iteration, a partial solution, which consists of only a start node at the first iteration or a connected set of nodes at any iteration afterward, is expanded by uniformly selecting at random a node neighboring the partial solution, until  $k$  nodes are included. We leverage the notion of Optimal Computing Budget Allocation (OCBA) [3] to randomly generate

**Table 1: Parameter Summary**

Notation	Description	Impact
$\tau_{i,j}$	social tightness score between node $v_i$ and $v_j$	set to be a negative value if $v_i$ and $v_j$ are foes
$\eta_i$	interest score of node $v_i$	set to be a negative value if $v_i$ does not like the topic
$\lambda_i$	weighting between interest score and tightness score of node $v_i$	set to be zero if $v_i$ only considers social tightness and one if $v_i$ only concerns topic interest
$T$	total computation budget	trade-off between solution quality and computation time
$m$	number of start nodes	sampling coverage

more final solutions from each start node that has more potential to generate the final solutions with high willingness. Later we will prove that the number of final solutions generated from each start node is optimally assigned.

After this, we enhance *CBAS* to *CBAS-ND* by differentiating the selection of the nodes neighboring each partial solution. During each iteration of *CBAS*, each neighboring node is treated equally and chosen uniformly at random. A simple way to improve *CBAS* is to associate each neighboring node with a different probability according to its interest score and social tightness scores of incident edges. Yet, this assignment is similar to the greedy algorithm insofar as it limits the scope to only the local information associated with each node thereby making it difficult to generate a final solution with high willingness. To prevent the generation of only a local optimal solution, *CBAS-ND* deploys the cross entropy method according to results at the previous stages in order to optimally assign a probability to each neighboring node.

One advantage of the proposed randomized algorithms is that the tradeoff between the solution quality and execution time can be easily controlled by assigning different  $T$ , which denotes the number of randomly generated final solutions. Under a given  $T$ , if  $m$  start nodes are generated, the above algorithms can optimally divide  $T$  into  $m$  parts for the  $m$  start nodes to find final solutions with high willingness. Moreover, we prove that *CBAS* is able to find a solution with an approximation ratio. Compared with *CBAS*, we further prove that the solution quality of *CBAS-ND* is better under the same computation budget  $T^8$ . The detailed settings of  $T$  and  $m$  will be analyzed in Section 5. In addition, the parameter summary with their impacts on the solution are shown in Table 1.

In the following, we first present *CBAS* to optimally allocate the computational budgets to different start nodes (Section 3.1) and then derive the approximation ratio in Section 3.2. Algorithm *CBAS-ND* will be presented in Section 4.

### 3.1 Allocation of Computational Budget for Start Nodes

Given the total computational budgets  $T$  specified by users, a simple approach first randomly selects  $m$  start nodes and then expands each start node to  $\frac{T}{m}$  final solutions. However, this homogeneous approach does not give priority to the start nodes that have more potential to generate final solutions with high willingness.

<sup>8</sup>It is worth noting that randomization is performed only in expanding a start node to a final solution, not in the selection of a start node. This is because the approximation ratio is not able to be achieved if a start node is decided randomly.

In contrast, *CBAS* optimally allocates more resources to the start nodes with high willingness with the following phases.

1. *Selection and Evaluation of Start Nodes*: This phase first selects  $m$  start nodes according to the interest scores and social tightness scores. Afterward, each start node is randomly expanded to a few final solutions. We iteratively select and add a neighboring node uniformly at random to a partial solution, until  $k$  nodes are selected. The willingness of each final solution is evaluated for the next phase to allocate different computational budgets to different start nodes.
2. *Allocation of Computational Budgets*: This phase derives the computational resources optimally allocated to each start node according to the previous sampled willingness.

To optimally allocate the computational budgets for each start node, we first define the solution quality as follows.

*Definition 1.* The solution quality, denoted by  $Q$ , is defined as the maximum willingness among all maximal sampled results of the  $m$  start nodes,

$$Q = \max\{J_1^*, J_2^*, \dots, J_i^*, \dots, J_m^*\},$$

where  $J_i^*$  is a random variable representing the maximal willingness sampled from a final solution expanded from start node  $v_i$ .

Since the maximal sampled result  $J_i^*$  of start node  $v_i$  is related to the number of sampling times  $N_i$ , i.e., the number of final solutions randomly generated from  $v_i$ , the mathematical formulation to optimize the computational budget allocation is defined as

$$\begin{aligned} & \max_{N_1, N_2, \dots, N_m} Q, \\ & \text{s.t. } N_1 + N_2 + \dots + N_m = T. \end{aligned}$$

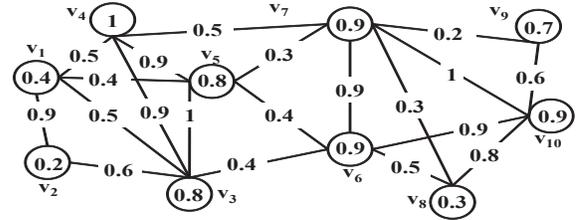
Let  $v_b$  denote the start node that are able to generate the solution with the highest willingness. Obviously, the optimal solution in the above maximization problem is to allocate all the computational budgets to  $v_b$ . However, since  $v_b$  is not given *a priori*, *CBAS* divides the resource allocation into  $r$  stages, and each stage adjusts the allocation of computational budgets  $\frac{T}{r}$  to different start nodes according to the sampled willingness from the partial solutions in previous stages.

For each node, phase 1 of *CBAS* first adds the interest score and the social tightness scores of incident edges and then chooses the  $m$  nodes with the largest sums as the  $m$  start nodes. On the other hand, allocating more computational budgets to the start node with a larger sum, similar to the greedy algorithm, does not tend to generate a final solution with high willingness. For this reason, phase 2 evaluates the sampled willingness to allocate different computational budgets to each start node.

In stage  $t$  of phase 2, let  $N_{i,t}$  denote the computational budgets allocated to start node  $v_i$  at the  $t$ -th stage. The ratio of computational budgets  $N_{i,t}$  and  $N_{j,t}$  allocated to any two start nodes  $v_i$  and  $v_j$  is

$$\frac{N_{i,t}}{N_{j,t}} = \left( \frac{d_i - c_b}{d_j - c_b} \right)^{N_b},$$

where  $d_i$  denotes the best sampled willingness of the partial solutions expanded from start node  $v_i$  in the previous stages  $1, \dots, t-1$ . Notice that  $v_b$  here is the start node that enjoys the highest willingness sampled in the previous stages,  $N_b$  is the overall computational budgets allocated to  $v_b$  in the previous stages, and  $c_b$  denotes the worst sampled willingness of the partial solution expanded from start node  $v_b$  in the previous stages. Later, we will prove that the above budgets allocation in each stage is optimal. However, if the allocated computational budgets for a start node is 0 at the  $t$ -th stage, we prune off the start node in the following  $(t+1)$ -th stage.



**Figure 2: An illustrative example for CBAS and CBAS-ND**

**EXAMPLE 1.** Figure 2 presents an illustrative example for *CBAS* with  $n = 10$ ,  $k = 5$ , and  $m = 2$ . Phase 1 first chooses two start nodes by summing up the topic interest score and the social tightness scores for every node. Therefore,  $v_3$  with  $0.8 + 0.6 + 0.5 + 0.9 + 1 + 0.4 = 4.2$  and  $v_{10}$  with  $0.9 + 0.6 + 1 + 0.9 + 0.8 = 4.2$  are selected. Next, let  $T = 20$ ,  $P_b = 0.7$  and  $\alpha = 0.9$  in this example, and the number of stages is thus  $r \leq \frac{Tk \ln \alpha}{n \ln(\frac{2(1-P_b)}{k-1})} = \frac{20 \cdot 5 \ln 0.9}{10 \ln(0.6)} \approx$

2. Each start node generates 5 samples at the first stage. In the beginning, the node selection probability of start node  $v_3$ , i.e.,  $\vec{p}_{3,1}$ , is set to be  $(\frac{4}{9}, \frac{4}{9}, 1, \frac{4}{9}, \frac{4}{9}, \frac{4}{9}, \frac{4}{9}, \frac{4}{9}, \frac{4}{9}, \frac{4}{9})$ . The intermediate solution obtained so far is denoted as  $V_S$ , and the candidate attendees extracted so far is denoted as  $V_A$ . Therefore, the total willingness of  $V_S = \{v_3\}$  is 0.8, and  $V_A = \{v_1, v_2, v_4, v_5, v_6\}$ . Since the node selection probability is homogeneous in the first stage, we randomly select  $v_6$  from  $V_A$  to expand  $V_S$ . Now the total willingness of  $V_S = \{v_3, v_6\}$  is  $W(V_S) = 0.8 + 0.4 + 0.9 = 2.1$ , and  $V_A = \{v_1, v_2, v_4, v_5, v_7, v_8, v_{10}\}$ . The process of expanding  $V_S$  continues until the cardinality of  $V_S$  reaches 5, and we record the first sample result  $X_{3,1} = \langle 1, 0, 1, 1, 1, 1, 0, 0, 0, 0 \rangle$  with the total willingness 8.9, the worst result of  $v_3$  ( $c_3 = 8.9$ ), and the best result of  $v_3$  ( $d_3 = 8.9$ ). The other sample results from start node  $v_3$  are  $X_{3,2} = \langle 1, 1, 1, 1, 1, 0, 0, 0, 0, 0 \rangle$  with the total willingness 8.9,  $X_{3,3} = \langle 0, 1, 1, 0, 1, 1, 0, 1, 0, 0 \rangle$  with the total willingness 5.9,  $X_{3,4} = \langle 0, 1, 1, 1, 1, 0, 1, 0, 0, 0 \rangle$  with the total willingness 7.9, and  $X_{3,5} = \langle 0, 0, 1, 0, 1, 1, 1, 0, 0, 1 \rangle$  with the total willingness 9.2. The worst and the best results of  $v_3$  are updated to  $c_3 = 5.9$  and  $d_3 = 9.2$ , respectively. After sampling from node  $v_3$ , we repeat the above process for start node  $v_{10}$ . The worst result is  $c_{10} = 6.9$ , and the best result is  $d_{10} = 8.9$ .

To allocate the computational budgets for the second stage, i.e.,  $r = 2$ , we first find the allocation ratio  $r_3 : r_{10} = 1 : (\frac{8.8-5.9}{9.2-5.9})^5 = 1 : 0.524$ . Therefore, the allocated computational budgets for start nodes  $v_3$  and  $v_{10}$  are  $\frac{10}{1.524} \approx 7$  and  $\frac{5.24}{1.524} \approx 3$ , respectively. At the second stage, the best results of  $v_3$  and  $v_{10}$  are 9.2 and 8.9, respectively. Finally, we obtain the solution  $\{v_3, v_5, v_6, v_7, v_{10}\}$  with the total willingness 9.2.

### 3.2 Theoretical Result of CBAS

To correctly allocate the computational budgets  $T$  to  $m$  start nodes, we first derive the optimal ratio of computational budgets for any two start nodes. Afterward, we find the probability  $P_b$  that node  $v_b$  is actually the start node which is able to generate the highest willingness in each stage. Finally, we derive the approximation ratio and analyze the complexity of *CBAS*.

*Definition 2.* A random variable, denoted as  $J_i$ , is defined to be the sampled value in start node  $v_i$ .

The literature of OCBA indicates that the distribution of random variable  $J_i$  in most applications is a normal distribution, but the allocation results are very close to the one with the uniform distribution [3, 6]. Therefore, given space constraints,  $J_i$  here is first handled as the uniform distribution in  $[c_i, d_i]$ , and the derivation for

the normal distribution is presented in [18]. The probability density function and cumulative distribution function are formulated as

$$p_{J_i}(x) = \begin{cases} \frac{1}{d_i - c_i} & \text{if } c_i \leq x \leq d_i \\ 0 & \text{otherwise.} \end{cases}$$

$$P_{J_i}(x) = \begin{cases} 0 & \text{if } x \leq c_i. \\ \frac{x - c_i}{d_i - c_i} & \text{if } c_i \leq x \leq d_i. \\ 1 & \text{otherwise.} \end{cases}$$

Therefore, the probability density function of the maximal value  $J_i^*$  is  $N_i P_{J_i}(x)^{N_i-1} p_{J_i}(x)$ . Moreover, the cumulative density function of the maximal value  $J_i^*$  is  $P_{J_i}(x)^{N_i}$ . As proved in [18], given the best start node  $v_b$ , the probability that  $J_i^*$  exceeds  $J_b^*$  is at most  $\frac{1}{2}(\frac{d_i - c_b}{d_b - c_b})^{N_b}$ , i.e.,

$$p(J_b^* \leq J_i^*) \leq \frac{1}{2}(\frac{d_i - c_b}{d_b - c_b})^{N_b} \quad (2)$$

With the result above, we allocate the computational budgets by

$$\frac{N_i}{N_j} = \frac{P(J_i^* \geq J_b^*)}{P(J_j^* \geq J_b^*)} = (\frac{d_i - c_b}{d_j - c_b})^{N_b}. \quad (3)$$

Since it is impossible to enumerate every final solution expanded from a start node, the ratio of the computational budget allocation is optimal in OCBA [3] if the first equality in Eq. (3) holds. Thus, it is optimal to allocate the computational budgets to  $N_i$  and  $N_j$  according to the ratio  $(\frac{d_i - c_b}{d_j - c_b})^{N_b}$ . Notice that if  $d_i$  is smaller than  $c_b$ , the probability that  $J_b^*$  is smaller than  $J_i^*$  is zero.

Intuitively, the above result indicates that if the best random sample, i.e.,  $d_i$ , from a start node is small, it is unnecessary to repeat the sampling process too many times since the users nearby the start node are not really interested in the activity or they have an estranged friendship. On the other hand, as the number of sample times increases, it is expected that the identified best start node enjoys the highest willingness.

We first analyze the probability  $P_b$  that  $v_b$ , as decided according to the samples in the previous stages, is actually the start node that generates the highest willingness. Let  $\alpha$  denote the closeness ratio between the maximum of the start node with the highest willingness and the maximum of other start nodes, i.e.,  $\alpha = (d_a - c_b)/(d_b - c_b)$ , where  $v_a$  generates the maximum willingness among other start nodes. For WASO with parameter  $(m, T)$ , where  $m$  is the number of start nodes and  $T$  is the total computational budgets, the probability  $P_b$  that  $v_b$  selected according to the previous stages is actually the start node with the highest willingness is at least  $1 - \frac{1}{2}(m-1)\alpha^{\frac{T}{m}}$  as proved in [18].

Given the total budgets  $T$ , the following theorem derives a lower bound of the solution obtained by CBAS.

**THEOREM 3.** *For a WASO optimization problem with  $r$ -stage computational budget allocation, the maximum willingness  $E[Q]$  from the solution of CBAS is at least  $N_b(\frac{1}{N_b+1})^{\frac{N_b+1}{N_b}} \cdot Q^*$ , where  $N_b$  after  $r$  stages is  $\frac{4+m(r-1)}{4rm}T$ , and  $Q^*$  is the optimal solution.*

**PROOF.** We first derive the lower bound of  $E[Q]$  as follows. The random variable  $Q$  is denoted as  $\max\{J_1^*, \dots, J_m^*\}$ . The cumulative density function is

$$\begin{aligned} F_Q(Q \leq \Delta) &= F(\max\{J_1^*, \dots, J_m^*\} \leq \Delta) \\ &= F(J_1^* \leq \Delta, J_2^* \leq \Delta, \dots, J_m^* \leq \Delta) \\ &= (\frac{\Delta - c_1}{d_1 - c_1})^{N_1} (\frac{\Delta - c_2}{d_2 - c_2})^{N_2} \dots (\frac{\Delta - c_m}{d_m - c_m})^{N_m}, \end{aligned}$$

where  $F_{J_i^*}(\Delta) = 1$ , for  $\Delta \geq d_i$ . After exploiting Markov's inequality,

$$F_Q(Q \geq \Delta) \leq \frac{E[Q]}{\Delta}.$$

$$E[Q] \geq \Delta(1 - (\frac{\Delta - c_b}{d_b - c_b})^{N_b}).$$

We normalize the lower bound and upper bound with  $c_b = 0$  and  $d_b = 1$ . Let  $\Delta$  be the top- $\rho$  percentile solution value, i.e.  $\Delta = c_b + (1 - \rho)(d_b - c_b)$ . Therefore,

$$E[\tilde{Q}] \geq (1 - \rho)(1 - (1 - \rho)^{N_b}).$$

To find the maximum  $(1 - \rho)(1 - (1 - \rho)^{N_b})$ , we let

$$\frac{\partial(1 - \rho)(1 - (1 - \rho)^{N_b})}{\partial \rho} = 0.$$

Therefore,

$$E[Q] \geq N_b(\frac{1}{N_b+1})^{\frac{N_b+1}{N_b}} \cdot Q^*.$$

If the computational budget allocation is  $r$ -stages with  $T \geq mr \frac{\ln(m-1)}{\ln(\frac{1}{\alpha})}$ ,  $N_b$  is  $\frac{T}{r}/m + \frac{1}{2} \frac{r-1}{2r} T$ , which is  $\frac{4+m(r-1)}{4rm}T$ .  $\square$

**Time Complexity of CBAS.** The time complexity of CBAS contains two parts. The first phase selects  $m$  start nodes with  $O(E + n + m \log n)$  time, where  $O(E)$  is to sum up the interest and social tightness scores,  $O(n + m \log n)$  is to build a heap and extract  $m$  nodes with the largest sum. Afterward, the second phase of CBAS includes  $r$  stages, and each stage allocates the computation resources with  $O(m)$  time and generates  $O(\frac{T}{r})$  new partial solutions with  $k$  nodes for all start nodes. Therefore, the time complexity of the second phase is  $O(r(m + \frac{T}{r})) = O(kT)$ , and CBAS therefore needs  $O(E + m \log n + kT)$  running time.

## 4. NEIGHBOR DIFFERENTIATION IN RANDOMIZATION

### 4.1 Greedy Neighbor Differentiation

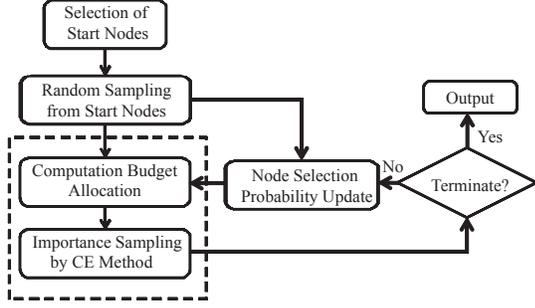
In Section 3.1, CBAS includes two phases. The first phase initiates the start nodes, while the second phase allocates different computational budgets to each start node to generate different numbers of final solutions. During the growth of a partial solution, CBAS chooses a neighboring node uniformly at random at each iteration. In other words, each neighboring node of the partial solution is treated equally. It is expected that this homogeneous strategy needs more computational budgets, because a neighboring node inclined to generate a final solution with high willingness is not associated with a higher probability.

To remedy this issue, a simple algorithm *RGreedy* (randomized greedy) associates each neighboring node with a different probability according to its interest score and social tightness scores of the edges incident to the partial solution  $S_{t-1}$  obtained in the previous stage, which is similar to the concept in the greedy algorithm. Given  $S_{t-1}$ , the ratio of the probabilities that *RGreedy* selects nodes  $v_i$  and  $v_j$  at iteration  $t$  is

$$\frac{P(v_i|S_{t-1})}{P(v_j|S_{t-1})} = \frac{W(\{v_i\} \cup S_{t-1})}{W(\{v_j\} \cup S_{t-1})},$$

where  $W(\{v_i\} \cup S_{t-1})$  denotes the willingness of the node set  $\{\{v_i\} \cup S_{t-1}\}$ . At each iteration, *RGreedy* randomly selects a vertex in accordance with  $W(\{v_j\} \cup S_{t-1})$ , until  $k$  nodes are included.

Intuitively, *RGreedy* can be regarded as a *randomized version* of the greedy algorithm with  $m$  start nodes, while the greedy algorithm is a deterministic algorithm with only one start node. Thus, similar to the greedy algorithm, the assignment of the probability limits the scope to only the local information associated with



**Figure 3: Flowchart of CBAS-ND**

each node and incident edges. It is envisaged that *RGreedy* is difficult to generate a final solution with high willingness, which is also demonstrated in Section 5. In contrast, we propose *CBAS-ND* by exploiting the cross entropy method according to the sampling partial solutions in previous stages, in order to optimally assign a probability to each neighboring node.

## 4.2 Neighbor Differentiation with Cross Entropy

We enhance *CBAS* to *CBAS-ND* to differentiate the selection of a node neighboring each partial solution. Algorithm *CBAS* is divided into  $r$  stages. In each stage, it optimally adjusts the computational budgets allocated to each start node according to the sampled maximum and minimum willingness in previous stages. To effectively improve *CBAS*, *CBAS-ND* takes advantage of the cross entropy method [17] to achieve importance sampling by adaptively assigning a different probability to each neighboring node from the sampled results in previous stages. In contrast to *RGreedy* with a greedy-based probability vector assigned to the neighboring nodes, it is expected that *CBAS-ND* is able to obtain final solutions with better quality. Indeed, later in Section 4.3, we prove that the solution quality of *CBAS-ND* is better than *CBAS* with the same computation budget  $T$ .

The flowchart of *CBAS-ND* is shown in Figure 3. We first define the node selection probability vector in *CBAS-ND*, which specifies the probability to add a node in  $G$  to the current partial solution expanded from a start node.

*Definition 3.* Let  $\vec{p}_{i,t}$  denote the node selection probability vector for start node  $v_i$  in stage  $t$ .

$$\vec{p}_{i,t} = \langle p_{i,t,1}, \dots, p_{i,t,j}, \dots, p_{i,t,n} \rangle,$$

where  $p_{i,t,j}$  is the probability of selecting node  $v_j$  for start node  $v_i$  in the  $t$ -th stage.

In the first stage, the node selection probability vector  $\vec{p}_{i,1}$  for each start node  $v_i$  is initialized homogeneously for every node, i.e.  $\vec{p}_{i,1,j} = (k-1)/|V|, \forall v_j \in G, v_j \neq v_i$ . That is, computational budgets  $\frac{T}{m}$  are identically assigned to each start node, and the probability associated with every node is also the same. However, different from *CBAS* and *RGreedy*, *CBAS-ND* here examines the top- $\rho$  samples for each start node  $v_i$  to generate  $\vec{p}_{i,2}$ , so that the node probability will be differentiated according to sampled result in stage 1.

*Definition 4.* A Bernoulli sample vector which is denoted as  $X_{i,q} = \langle x_{i,q,1}, \dots, x_{i,q,j}, \dots, x_{i,q,n} \rangle$ , is defined to be the  $q$ -th sample vector from start node  $v_i$ , where  $x_{i,q,j}$  is 1 if node  $v_j$  is selected in the  $q$ -th sample and 0 otherwise.

*Definition 5.*  $\gamma_{i,t}$  is denoted as the top- $\rho$  sample quantile of the performances in the  $t$ -th stage of start node  $v_i$ , i.e.,  $\gamma_{i,t} = W_{(\lceil \rho N_{i,t} \rceil)}$ .

Specifically, after collecting  $N_{i,1}$  samples  $X_{i,1}, X_{i,2}, \dots, X_{i,q}, \dots, X_{i,N_{i,1}}$  generated from  $\vec{p}_{i,1}$  for start node  $v_i$ , Node Selection Probability Update in Figure 3 calculates the total willingness  $W(X_{i,q})$  for each sample, and sorts them in the descending order,  $W_{(1)} \geq \dots \geq W_{(N_{i,1})}$ , while  $\gamma_{i,1}$  denotes the willingness of the top- $\rho$  performance sample, i.e.  $\gamma_{i,1} = W_{(\lceil \rho N_{i,1} \rceil)}$ . With those sampled results, the selection probability  $p_{i,2,j}$  of every node  $v_j$  in the second stage is derived according to the following equation,

$$p_{i,t+1,j} = \frac{\sum_{q=1}^{N_{i,t}} I_{\{W(X_{i,q}) \geq \gamma_{i,t}\}} x_{i,q,j}}{\sum_{q=1}^{N_{i,t}} I_{\{W(X_{i,q}) \geq \gamma_{i,t}\}}}, \quad (4)$$

where the indicator function  $I_{\{W(X_{i,q}) \geq \gamma_{i,t}\}}$  is defined on the feasible solution space  $\chi$  such that  $I_{\{W(X_{i,q}) \geq \gamma_{i,t}\}}$  is 1 if the willingness of sample  $X_{i,q}$  exceeds a threshold  $\gamma_{i,t} \in \mathbb{R}$ , and 0 otherwise. Eq. (4) derives the node selection probability vector by fitting the distribution of top- $\rho$  performance samples. Intuitively, if node  $v_j$  is included in most top- $\rho$  performance samples in  $t$ -th stage,  $p_{i,t+1,j}$  will approach 1 and be selected in  $(t+1)$ -th stage.

Later in Section 4.3, we prove that the above probability assignment scheme is optimal from the perspective of cross entropy. Eq. (4) minimizes the Kullback-Leibler cross entropy (KL) distance [17] between node selection probability  $\vec{p}_{i,t}$  and the distribution of top- $\rho$  performance samples, such that the performance of random samples in  $t+1$  is guaranteed to be closest to the top- $\rho$  performance samples in  $t$ . Therefore, by picking the top- $\rho$  performance samples to generate the partial solutions in the next stage, the performance of random samples is expected to be improved after multiple stages. Most importantly, by minimizing the KL distance, the convergence rate is maximized.

Moreover, it is worth noting that a smoothing technique is necessary to be included in adjusting the selection probability vector,

$$\vec{p}_{i,t+1} = w \vec{p}_{i,t+1} + (1-w) \vec{p}_{i,t},$$

to avoid setting 0 or 1 in the selection probability for any node  $v_j$ , because  $v_j$  will no longer appear or always appear in this case. An example illustrating *CBAS-ND* is provided as follows. As demonstrated in Section 4.3, the solution quality of *CBAS-ND* is better than *CBAS* with the same computation budget.

*EXAMPLE 2.* Take Figure 2 as an illustrating example of *CBAS-ND*. Since *CBAS-ND* is different from *CBAS* in the second phase to obtain the node selection probability vector, we continue from the result of the first phase in Section 3, i.e., the allocated computational budgets for start node  $v_3$  and  $v_{10}$  are 7 and 3 respectively, and illustrate the second phase of *CBAS-ND* with Figure 2.

By sorting the willingness samples  $X_3$  to  $W = \langle 9.2, 8.9, 8.9, 7.9, 5.9 \rangle$ ,  $\gamma_1$  is equal to  $W_{(\lceil \frac{1}{2} 5 \rceil)} = 8.9$ . Therefore, the samples with the total willingness exceeding 8.9 include  $X_1, X_2$ , and  $X_5$ , which are used to update the node selection probability  $\vec{p}_{3,2}$  to  $\langle \frac{2}{3}, \frac{1}{3}, 1, \frac{2}{3}, 1, \frac{2}{3}, \frac{1}{3}, 0, 0, 0 \rangle$ . Then, the smoothing technique is adopted with  $w = 0.6$ , and the node selection probability  $\vec{p}_{3,2}$  becomes

$$\begin{aligned} \hat{p}_{3,2} &= 0.6 \langle \frac{2}{3}, \frac{1}{3}, 1, \frac{2}{3}, 1, \frac{2}{3}, \frac{1}{3}, 0, 0, 0 \rangle \\ &\quad + 0.4 \langle \frac{4}{9}, \frac{4}{9}, 1, \frac{4}{9}, \frac{4}{9}, \frac{4}{9}, \frac{4}{9}, \frac{4}{9}, \frac{4}{9} \rangle \\ &= \langle \frac{5.2}{9}, \frac{3.4}{9}, 1, \frac{5.2}{9}, \frac{7}{9}, \frac{5.2}{9}, \frac{3.4}{9}, \frac{1.6}{9}, \frac{1.6}{9}, \frac{1.6}{9} \rangle. \end{aligned}$$

After sampling from node  $v_3$ , we repeat the above process for start node  $v_{10}$ . The worst result is  $c_{10} = 6.9$ , the best result is  $d_{10} = 8.9$ , and the node selection probability is  $\vec{p}_{10,2} = \langle \frac{1.6}{9}, \frac{1.6}{9}, \frac{1.6}{9}, \frac{3.4}{9}, \frac{5.2}{9}, \frac{5.2}{9}, \frac{7}{9}, \frac{5.2}{9}, \frac{5.2}{9}, 1 \rangle$ . At the second stage, the best results of  $v_3$

and  $v_{10}$  are 9.7 and 8.9, respectively. Finally, we obtain the solution  $\{v_3, v_4, v_5, v_6, v_7\}$  with the total willingness 9.7, which is also the optimal solution in this example and outperforms the solution obtained from CBAS.

### 4.3 Theoretical Result of CBAS-ND

In the following, we prove that the probability assignment with the cross-entropy method [17] in Eq. (4) is optimal. The idea of cross-entropy method originates from importance sampling<sup>9</sup>, i.e., by changing the distribution of sampling on different neighbors such that the neighbors having the potential to boost the willingness are able to be identified and included. Therefore, we first derive the probability of a random sample according to the sampling results in previous stages. After this, we introduce importance sampling and derive the node selection probability vector in the WASO problem to replace the original sampling vector such that the Kullback-Leibler cross entropy (KL) distance between the sampling vector and the optimal importance sampling vector is minimized. Intuitively, a small KL distance ensures that two distributions are very close and implies that the node selection probability vector is optimal because the KL distance between the node selection probability vector in *CBAS-ND* and optimal node selection probability vector is minimized. Equipped with importance sampling vector, later in this section we prove that the expected solution quality of *CBAS-ND* is better than *CBAS*.

More specifically, let  $\chi$  denote the feasible solution space, and  $X$  is a feasible solution in  $\chi$ , i.e.,  $X \in \chi$ . WASO chooses a group of attendee  $X^*$  to find the maximum willingness  $\gamma^*$ ,

$$W(X^*) = \gamma^* = \max_{X \in \chi} W(X).$$

To derive the probability that the willingness of a random sample  $X$  exceeds a large value  $\gamma$ , i.e.  $W(X) \geq \gamma$ , it is necessary for *CBAS* to generate many samples given that it uniformly selects a neighboring node at random. In contrast, *CBAS-ND* leverages the notion of importance sampling to change the distribution of sampling on different neighbors. In the following, we first derive the optimal distribution of sampling. First, for the initial partial solution with one start node, let  $f(X; \vec{p})$  denote the probability density function of generating a sample  $X$  according a real-valued vector  $\vec{p}$ , and  $f(\cdot; \vec{p})$  is a family of probability density functions on  $\chi$ , i.e.,

$$f(\cdot; \vec{p}) = \{f(X; \vec{p}) | X \in \chi\}.$$

*CBAS* can be regarded as a special case of *CBAS-ND* with the homogeneous assignment on the above vector. A random sample  $X(\vec{p})$  for  $\vec{p} = \{p_1, \dots, p_j, \dots, p_n\}$  is generated with probability  $f(X(\vec{p}); \vec{p})$ , where  $p_j$  denotes the probability of selecting node  $v_j$  and is the same for all  $j$  in *CBAS*. The probability  $P_{\vec{p}}(\gamma)$  that the willingness of  $X(\vec{p})$  exceeds the threshold  $\gamma$  is

$$\begin{aligned} P_{\vec{p}}(\gamma) &= \mathbb{P}_{\vec{p}}(W(X(\vec{p}))) \geq \gamma) \\ &= \sum_{X \in \chi} I_{\{W(X(\vec{p})) \geq \gamma\}} f(X(\vec{p}); \vec{p}). \end{aligned}$$

However, the above equation is impractical and inefficient for a large solution space, because it is necessary to scan the whole solution space  $\chi$  and sum up the probability  $f(X(\vec{p}); \vec{p})$  of every sample  $X$  with  $W(X(\vec{p})) \geq \gamma$ . To more efficiently address this issue, a direct way to derive the estimator  $\hat{P}_{\vec{p}}(\gamma)$  of  $P_{\vec{p}}(\gamma)$  is by

<sup>9</sup>Importance sampling [17] is used to estimate the properties of a target distribution by using the observations from a different distribution. By changing the distribution, the "important" values can be effectively extracted and emphasized by sampling more frequently to reduce the sample variance.

employing a crude Monte-Carlo simulation and drawing  $N$  random samples  $X_1(\vec{p}), \dots, X_N(\vec{p})$  by  $f(\cdot, \vec{p})$  to find  $P_{\vec{p}}(\gamma)$ ,

$$\hat{P}_{\vec{p}}(\gamma) = \frac{1}{N} \sum_{i=1}^N I_{\{W(X_i(\vec{p})) \geq \gamma\}}.$$

However, the crude Monte-Carlo simulation poses a serious problem when  $\{W(X(\vec{p})) \geq \gamma\}$  is a rare event since rare events are difficult to be sampled, and thus a large sample number  $N$  is necessary to estimate  $P_{\vec{p}}(\gamma)$  correctly.

Based on the above observations, *CBAS-ND* attempts to find the distribution  $f(X(\vec{p}); \vec{p})$  based on another importance sampling pdf  $f(X(\vec{p}_g^*); \vec{p}_g^*)$  to reduce the required sample number. For instance, consider a network with 3 nodes, i.e.  $V = \{v_1, v_2, v_3\}$ , and the 2-node group where the maximum willingness  $\gamma^*$  is  $\{v_1, v_2\}$ . The expected number of samples with node selection vector  $\{\frac{2}{3}, \frac{2}{3}, \frac{2}{3}\}$  in *CBAS* is larger than the node selection vector of  $\{1, 1, 0\}$  in *CBAS-ND*. In finer detail, let  $X_i(\vec{p}_g^*)$  denote the  $i$ -th random sample generated by  $f(X(\vec{p}_g^*); \vec{p}_g^*)$ . *CBAS-ND* first creates random samples  $X_1(\vec{p}_g^*), \dots, X_N(\vec{p}_g^*)$  generated by  $\vec{p}_g^*$  on  $\chi$  and then estimates  $\hat{P}_{\vec{p}}(\gamma)$  according to the likelihood ratio (LR) estimator  $\frac{f(X_i(\vec{p}_g^*); \vec{p}_g^*)}{f(X_i(\vec{p}_g); \vec{p}_g)}$ ,

$$\begin{aligned} \hat{P}_{\vec{p}}(\gamma) &= \frac{1}{N} \sum_{i=1}^N I_{\{W(X_i(\vec{p})) \geq \gamma\}} \\ &= \frac{1}{N} \sum_{i=1}^N \{I_{\{W(X_i(\vec{p}_g^*)) \geq \gamma\}} \frac{f(X_i(\vec{p}_g^*); \vec{p}_g^*)}{f(X_i(\vec{p}_g); \vec{p}_g)}\}. \end{aligned} \quad (5)$$

Notice that the above equation holds when  $N$  is infinity, but in most cases  $N$  only needs to be sufficiently large in practical implementation [5]. Now the question becomes how to derive  $\vec{p}_g^*$  for importance sampling pdf  $f(X(\vec{p}_g^*); \vec{p}_g^*)$  to reduce the number of samples. The optimal importance sampling pdf  $f^*(X_i(\vec{p}_g^*); \vec{p}_g^*)$  to correctly estimate  $P_{\vec{p}}(\gamma)$  thus becomes

$$f^*(X_i(\vec{p}_g^*); \vec{p}_g^*) = \frac{I_{\{W(X_i(\vec{p}_g^*)) \geq \gamma\}} f(X_i(\vec{p}_g^*); \vec{p}_g^*)}{P_{\vec{p}}(\gamma)}. \quad (6)$$

In other words, by substituting  $f(X_i(\vec{p}_g^*); \vec{p}_g^*)$  with  $f^*(X_i(\vec{p}_g^*); \vec{p}_g^*)$  in Eq. (5),  $\hat{P}_{\vec{p}}(\gamma) = \frac{1}{N} \sum_{i=1}^N P_{\vec{p}}(\gamma)$  holds, implying that only 1 sample is required to estimate the correct  $P_{\vec{p}}(\gamma)$ , i.e.,  $N = 1$ . However, it is difficult to find the optimal  $f^*(X(\vec{p}_g^*); \vec{p}_g^*)$  since it depends on  $P_{\vec{p}}(\gamma)$ , which is unknown *a priori* and is therefore not practical for WASO.

Based on the above observations, *CBAS-ND* optimally finds  $\vec{p}_g^*$  and the importance sampling pdf  $f(X(\vec{p}_g^*); \vec{p}_g^*)$  to minimize the Kullback-Leibler cross entropy (KL) distance between  $f(X(\vec{p}_g^*); \vec{p}_g^*)$  and optimal importance sampling pdf  $f^*(X(\vec{p}_g^*); \vec{p}_g^*)$ , where the KL distance measures two densities  $f^*$  and  $f$  as

$$D(f^*, f) = \sum_{X \in \chi} f^*(X) \ln f^*(X) - \sum_{X \in \chi} f^*(X) \ln f(X). \quad (7)$$

The first term in the above equation is related to  $f^*$  and is fixed, and minimizing  $D(f^*, f)$  is equivalent to maximizing the second term, i.e.,  $\sum_{X \in \chi} f^*(X) \ln f(X)$ . It is worth noting that the importance sampling pdf  $f(X(\vec{p}_g^*); \vec{p}_g^*)$  is referenced to a vector  $\vec{p}_g^*$ . Thus, after substituting  $f^*(X_i(\vec{p}_g^*); \vec{p}_g^*)$  in Eq. (6) into the Eq. (7), the reference vector  $\vec{p}_g^*$  of importance sampling pdf  $f(X(\vec{p}_g^*); \vec{p}_g^*)$  that maximizes the second term of Eq. (7) is the optimal reference vector  $\vec{p}_g^{*}$  with the minimum KL distance,

$$\vec{p}_g^* = \arg \max_{\vec{p}_g} \sum_{X \in \mathcal{X}} \frac{I_{\{W(X(\vec{p}_g)) \geq \gamma\}} f(X(\vec{p}_g); \vec{p}_g)}{P_{\vec{p}_g}(\gamma)} \ln f(X(\vec{p}_g); \vec{p}_g). \quad (8)$$

Since  $P_{\vec{p}_g}(\gamma)$  is not related to  $\vec{p}_g$ , Eq. (8) is equivalent to

$$\arg \max_{\vec{p}_g} \mathbb{E}_{\vec{p}_g} I_{\{W(X(\vec{p}_g)) \geq \gamma\}} \ln f(X(\vec{p}_g); \vec{p}_g),$$

Because it is computationally intensive to generate and compare every feasible  $\vec{p}_g$ , we estimate  $\mathbb{E}_{\vec{p}_g} I_{\{W(X(\vec{p}_g)) \geq \gamma\}}$  by drawing  $N$  samples as

$$\arg \max_{\vec{p}_g} \frac{1}{N} \sum_{i=1}^N I_{\{W(X_i(\vec{p}_g)) \geq \gamma\}} \ln f(X_i(\vec{p}_g); \vec{p}_g).$$

Specifically, *CBAS-ND* first generates random samples  $X_1, \dots, X_i, \dots, X_N$ , where  $X_i$  is the  $i$ -th sample and is a Bernoulli vector generated by a node selection probability vector  $\vec{p}_g$ , i.e.,  $X_i = (x_{i,1}, \dots, x_{i,n}) \sim \text{Ber}(\vec{p}_g)$ , where  $\vec{p}_g = \{p_1, \dots, p_j, \dots, p_n\}$  and  $p_j$  denotes the probability of selecting node  $v_j$ . The pdf  $f(X_i(\vec{p}_g); \vec{p}_g)$  is

$$f(X_i(\vec{p}_g); \vec{p}_g) = \prod_{j=1}^n p_j^{x_{i,j}} (1 - p_j)^{1 - x_{i,j}}.$$

To find the optimal reference vector  $\vec{p}^*$  with Eq. (8), we first calculate the first derivative w.r.t.  $p_j$ ,

$$\frac{\partial}{\partial p_j} \ln f(X_i(\vec{p}_g); \vec{p}_g) = \frac{\partial}{\partial p_j} \ln p_j^{x_{i,j}} (1 - p_j)^{1 - x_{i,j}}. \quad (9)$$

Since  $x_{i,j}$  can be either 0 or 1, Eq. (9) is simplified to

$$\frac{\partial}{\partial p_j} \ln f(X_i(\vec{p}_g); \vec{p}_g) = \frac{1}{(1 - p_j)p_j} (x_{i,j} - p_j).$$

The optimal reference vector  $\vec{p}^*$  is obtained by setting the first derivative of Eq. (8) to zero.

$$\begin{aligned} & \frac{\partial}{\partial p_j} \sum_{i=1}^N I_{\{W(X_i) \geq \gamma\}} \ln f(X_i(\vec{p}_g); \vec{p}_g) \\ &= \frac{1}{(1 - p_j)p_j} \sum_{i=1}^N I_{\{W(X_i) \geq \gamma\}} (x_{i,j} - p_j) = 0. \end{aligned}$$

Finally, the optimal  $p_j$  assigned to each node  $v_j$  is

$$p_j = \frac{\sum_{i=1}^N I_{\{W(X_i) \geq \gamma\}} x_{i,j}}{\sum_{i=1}^N I_{\{W(X_i) \geq \gamma\}}}.$$

We prove that *CBAS-ND* outperforms *CBAS* in solution quality in the extended version [18].

**THEOREM 4.** *The solution quality of CBAS-ND is better than CBAS under the same computation budget  $T$ .*

**Time Complexity of CBAS-ND.** *CBAS-ND* is different from *CBAS* in the second phase to find the node selection probability vector, which needs  $O(r(mn\rho\frac{T}{r} + \frac{T}{r}k)) = O(mn\rho T)$ . Therefore, the time complexity of *CBAS-ND* is  $O(E + m \log n + mn\rho T)$ . However, in reality we can directly set the probability to 0 for every node not neighboring a partial solution of a start node. Therefore, as shown in Section 5, the experimental result manifests that the execution time of *CBAS-ND* is not far from *CBAS*, and both *CBAS* and *CBAS-ND* are much faster than *RGreedy*.

## 4.4 Extension

In the following, we briefly discuss two basic extensions for dealing with the dynamic situation and improving the solution quality of *CBAS-ND*.

### 4.4.1 Online computation

In the process of social activity planning, some candidate attendees may not accept the invitations, and an online algorithm to adjust the solution according to user responses can help us handle the dynamic situation. If the online decision of multiple attendees are dependent, the situation is similar to the entangled transactions [11] in databases, in which it is necessary that transactions be processed coordinately in multiple entangled queries. Therefore, we extend *CBAS-ND* to cope with the dynamic situation as follows. If a user can not attend the activity, it is necessary to invite new attendees. Nevertheless, we have already sent invitation, and some of them have already confirmed to attend. Therefore, *CBAS-ND* regards those confirmed attendees as the initial solution in the second phase and removes the nodes that can not attend the activity from  $G$ . Therefore, the node selection probability vector  $\vec{p}_{i,t}$  will be updated to identify the new neighbors leading to better solutions according to the confirmed attendees. It is worth noting that the above online computation is fast since the start nodes in the first phase have been decided.

### 4.4.2 Backtracking

In addition to online computation, we extend *CBAS-ND* for backtracking to further improve the solution quality as follows. As shown in the previous work [5, 17], the criterion of convergence for Cross-Entropy method is that the node selection probability vector does not change over a number of iterations. Motivated by the above work, given the node selection probability  $\vec{p}_{i,t}$  of *CBAS-ND* at each stage  $t$ , we derive the difference  $z_i$  between  $\vec{p}_{i,t}$  and  $\vec{p}_{i,t-1}$  as follows.

$$z_i = \sum_j^n (\vec{p}_{i,t,j} - \vec{p}_{i,t-1,j})^2.$$

When the difference  $z_i$  between  $\vec{p}_{i,t}$  and  $\vec{p}_{i,t-1}$  is lower than a given threshold  $z_t$ , which indicates that the solution quality converges, we backtrack the solution by resetting the node selection probability  $\vec{p}_{i,t}$  to  $\vec{p}_{i,t-1}$  and re-sample.

### 4.4.3 CBAS-ND for Different Scenarios

For the scenarios of *couple and foe*, *invitation*, and *exhibition*, *CBAS-ND* can be directly applied by modifying the node and edge weights of the graph. For the scenario of *separate groups*, the start nodes are selected first, and the virtual node  $\vec{v}$  is then added to the selection set  $V_S$  to relax the connectivity constraint.

## 5. EXPERIMENTAL RESULTS

In this section, we first present the experiment setup in Section 5.1. The results of user study are provided in Section 5.2 and the performance of the proposed algorithms with different parameter settings on real datasets are evaluated in Section 5.3.

### 5.1 Experiment Setup

We implement *CBAS-ND* in Facebook and invite 137 people from various communities, e.g., schools, government, technology companies, and businesses to join our user study, to compare the solution quality and the time to answer WASO with manual coordination and *CBAS-ND* for demonstrating the need of an automatic group recommendation service. Each user is asked to plan 10 social activities with the social graphs extracted from their social networks in Facebook. The interest scores follow the power-law distribution according to the recent analysis [4] on real datasets, which has found the power exponent  $\beta = 2.5$ . The social tightness score between two friends is derived according to the widely

adopted model based on the number of common friends that represent the proximity interaction [2]. Then, social tightness scores and interest scores are normalized. Nevertheless, after the scores are returned by the above renowned models, each user is still allowed to fine-tune the two scores by themselves. The 10 problems explore various network sizes and different numbers of attendees in two different scenarios. In the first 5 problems, the user needs to participate the group activity and is inclined to choose her close friends, while the following 5 problems allow the user to choose an arbitrary group of people with high willingness. In other words, *CBAS-ND* in the first 5 problems always chooses the user as a start node. In addition to the user study, three real datasets are tested in the experiment. The first dataset is crawled from Facebook with 90, 269 users in the New Orleans network<sup>10</sup>. The other datasets are DBLP and Flickr datasets.<sup>11</sup> However, due to the space constraint, detailed experimental results of the DBLP and Flickr datasets are presented in the extended version [18].

In the following, we compare *DGreedy*, *RGreedy*, *CBAS*, *CBAS-ND*, and *IP* (*Integer Programming*) solved by IBM CPLEX in an HP DL580 server with four Intel E7-4870 2.4 GHz CPUs and 128 GB RAM. IBM CPLEX is regarded as the fastest general-purpose parallel optimizer, and we adopt it to solve the Integer Programming formulation for finding the optimal solution to WASO<sup>12</sup>. The details of Integer Programming formulation is presented in the extended version [18]. It is worth noting that even though *RGreedy* performs much better than its counterpart *DGreedy* and is closer to *CBAS* and *CBAS-ND*, it is computation intensive and not scalable to support a large group size. Therefore, we can only plot a few results of *RGreedy* in some figures. The default  $m$  is set to be  $n/k$  since  $n/k$  different  $k$ -person groups can be partitioned from a network with size of  $n$ . With  $m$  equal  $n/k$ , the start nodes averagely cover the whole network. Nevertheless, the experimental analysis manifests that  $m$  can be set to be smaller than  $n/k$  in WASO since the way we select start nodes efficiently prunes the start nodes which do not generate good solutions. The computational budget of *CBAS-ND* is not wasted much since the start node that do not generate good solutions will be pruned after the first stage. The default cross-entropy parameters  $\rho$  and  $w$  are 0.3 and 0.9 respectively, and  $\alpha$  is 0.99 as recommended by the cross-entropy method [17]. The results with different settings of parameters will be presented. Since *CBAS* and *CBAS-ND* natively support parallelization, we also implemented them with OpenMP for parallelization, to demonstrate the gain in parallelization with more CPU cores.

## 5.2 User Study

The weights  $\lambda$  and  $(1-\lambda)$  in Section 2 for interest scores and social tightness scores are directly specified by the users according to their preferences, and Figure 4(a) shows that the range of the weight mostly spans from 0.37 to 0.66 with the average as 50.3, indicating that both social tightness and interest are crucial factors in activity planning. Figures 4(b)-(e) compare manual coordination and *CBAS-ND* in the user study. It is worth noting that we generate the ground truth of user study with *IP* solved by IBM CPLEX to evaluate the solution quality. Figures 4(b) and (c) present the solution quality and running time with different network sizes, where the expected number of attendees  $k$  is 7. The user must be included in the group for *Manual-i* and *CBAS-ND-i*, and in the other two cases the user can arbitrarily choose a group with high willingness. The result indicates that the solutions obtained by *CBAS-ND*

<sup>10</sup><http://socialnetworks.mpi-sws.org/data-wosn2009.html>.

<sup>11</sup><http://socialnetworks.mpi-sws.org/data-imc2007.html>.

<sup>12</sup>Note that because WASO is NP-Hard, it is only possible to find the optimal solutions to WASO with IBM CPLEX in small cases.

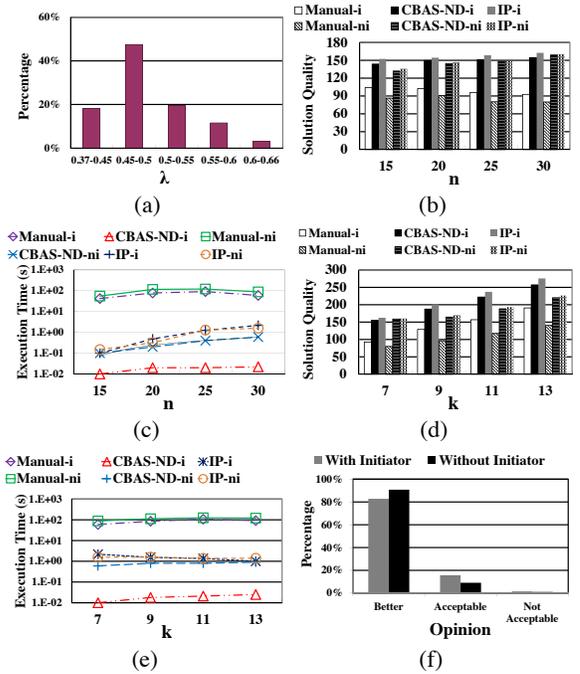


Figure 4: Results of user study

is very close to the optimal solutions acquired from solving *IP* with IBM CPLEX. WASO is challenging for manual coordination, even when the network contains only dozens of nodes. It is interesting that  $n = 30$  is too difficult for manual coordination because some users start to give up thus require smaller time for finding a solution. In addition, WASO is more difficult and more time-consuming in *Manual-ni* because it considers many more candidate groups.

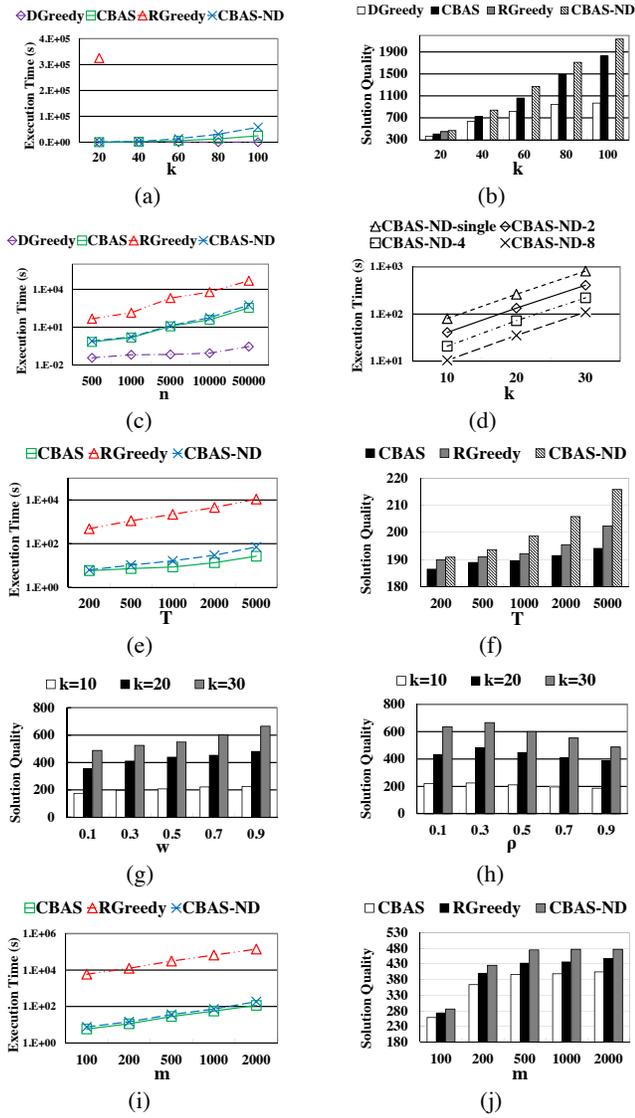
Figures 4(d) and (e) presents the results with different  $k$ . The results show that the solution quality obtained by manual coordination with  $k = 7$  is only 66% of *CBAS-ND*, since it is challenging for a person to jointly maximize the social tightness and interest. Similarly, we discover that some users start to give up when  $k = 13$ , and the processing time of manual selection grows when the user is not going to join the group activity. Finally, we return the solutions obtained by *CBAS-ND* to the users, and Figure 4(f) manifests that 98.5% of users think the solutions are better or acceptable, as compared to the solutions found by themselves. Therefore, it is desirable to deploy *CBAS-ND* as an automatic group recommendation service, especially to address the need of a large group in a massive social network nowadays.

## 5.3 Performance Comparison and Sensitivity Analysis

### 5.3.1 Experimental Results on Facebook

Figure 5(a) first presents the running time with different group sizes, i.e.,  $k$ . *RGreedy* is computationally intensive since it is necessary to sum up the interest scores and social tightness scores during the selection of a node neighboring each partial solution. Therefore, *RGreedy* is unable to return a solution within even 12 hours when the group size is larger than 20. In addition, the difference between *CBAS-ND* and *RGreedy* becomes more significant as  $k$  grows. Figure 5(b) presents the solution quality with different activity sizes, where  $m = \frac{n}{k}$ ,  $\rho = 0.3$ , and  $w = 0.9$ , respectively. The results indicate that *CBAS-ND* outperforms *DGreedy*, *RGreedy*, and *CBAS*, especially under a large  $k$ . The willingness of *CBAS-ND* is at least twice of the one from *DGreedy* when  $k = 100$ . On the other hand, *RGreedy* outperforms *DGreedy* since it has a chance to jump out of the local optimal solution.

In addition to the activity sizes, we compare the running time of *RGreedy*, *CBAS-ND*, and *DGreedy* with different social network

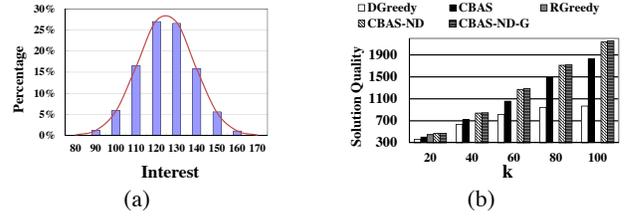


**Figure 5: Experimental results on Facebook dataset**

sizes in Figure 5(c) with  $k = 10$ . *DGreedy* is always the fastest one since it is a deterministic algorithm and generates only one final solution, but *CBAS* and *CBAS-ND* both require less than 10 seconds, whereas *RGreedy* requires more than  $10^3$  seconds. To evaluate the performance of *CBAS-ND* with multi-threaded processing, Figure 5(d) shows that we can accelerate the processing speed to around 7.6 times with 8 threads. The acceleration ratio is slightly lower than 8 because OpenMP forbids different threads to write at the same memory position at the same time. Therefore, it is expected that *CBAS-ND* with parallelization is promising to be deployed as a value-added *cloud service*.

Figures 5(e) and (f) compare the running time and solution quality of three randomized approaches under different total computational budgets, i.e.,  $T$ . As  $T$  increases, the solution quality of *CBAS-ND* increases faster than that of the others because it can optimally allocate the computation resources. The running time of *CBAS-ND* is slightly larger than that of *CBAS* since *CBAS-ND* needs to sort and extract the samples with high willingness in previous stages to generate better samples. Even though the solution quality of *RGreedy* is closer to *CBAS-ND* in some cases, both *CBAS* and *CBAS-ND* are faster than *RGreedy* by an order of  $10^{-2}$ .

Figure 5(g) presents the solution quality of *CBAS-ND* with different smoothing technique parameters, i.e.,  $w$ . Notice that the node selection probability vector is homogeneous if we set  $w$  to



**Figure 6: Experimental results of WASO with Gaussian distribution**

zero. The result shows that the best result is generated by  $w = 0.9$  for  $k = 10, 20$ , and  $30$ , implying that the convergence rate with  $w = 0.9$  is most suitable for WASO in the Facebook dataset. Figure 5(h) compares the top percentile of performance sample value  $\rho$ . The result manifests that the solution quality is not inversely proportional to  $\rho$ , because for a smaller  $\rho$ , the number of samples selected to generate the node selection probability vector decreases, such that the result converges faster to a solution.

Figures 5(i) and (j) present the running time and solution quality of *RGreedy*, *CBAS*, and *CBAS-ND* with different numbers of start nodes, i.e.,  $m$ . The results show that the solution quality in Figure 5(j) converges when  $m$  is equal to 500, which indicates that it is sufficient for  $m$  to be set as a value smaller than  $\frac{n}{k}$  as recommended by OCBA [3]. By assigning  $m = 500$  in the Facebook dataset, we can reduce the running time to only 20% of the running time in  $m = 2000$ , while the solution quality remains almost the same.

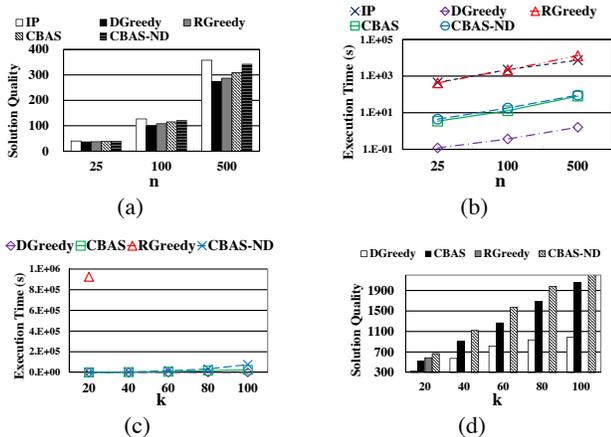
Figure 6(a) shows the interest histogram of random samples on Facebook, which indicates that the distribution follows a Gaussian distribution with the mean as 124.71 and variance as 13.83. The allocation ratio for the variant *CBAS-ND-G* of *CBAS-ND* by replacing the uniform distribution with the Gaussian distribution is derived in the extended version [18]. Figure 6(b) indicates that the solution quality of *CBAS-ND* and *CBAS-ND-G* is very close. In contrast to *CBAS-ND-G*, however, *CBAS-ND* is more efficient and easier to be implemented because it does not involve the probability integration to find the probability of the best start node.

### 5.3.2 Comparison with Integer Programming

To evaluate the solution quality of *CBAS-ND*, Figures 7(a) and (b) compare the solution quality and running time of *IP* (ground truth) with  $k = 10$ . Since WASO is NP-hard, i.e., the running time for obtaining the ground truth is unacceptably large, we extract 1000 small real datasets from the DBLP dataset with the node sizes as 25, 100, and 500 respectively. The result shows that the solution quality of *CBAS-ND* is very close to *IP*, while the running time is smaller by an order of  $10^{-2}$ . It is worth noting that *CBAS-ND* here is single-threaded, but *IP* is solved by IBM CPLEX (parallel version).

### 5.3.3 Experimental Results of WASO-dis

For separate groups, Figure 7(c) first presents the running time with different group sizes, i.e.,  $k$ , where  $m = \frac{n}{k}$ ,  $\rho = 0.3$ , and  $w = 0.9$ , respectively. For all algorithms, the virtual node  $\bar{v}$  is added to the selection set  $V_S$  to relax the connectivity constraint. *RGreedy* computes the incremental willingness of every node in  $V_A$  to the selection set  $V_S$ , where  $V_A$  includes all nodes, and thus are computationally intractable. Therefore, *RGreedy* is unable to return a solution within 24 hours when the group size is larger than 20. Figure 7(d) presents the solution quality with different activity sizes. The results indicate that *CBAS-ND* outperforms *DGreedy*, *RGreedy*, and *CBAS*, especially under a large  $k$ . In addition, compared to the experimental results in WASO, the difference between *CBAS-ND* and *DGreedy* becomes more significant as  $k$  increases. The reason is that the greedy algorithm selects the node with the largest incremental willingness to the current group and thus is in-



**Figure 7: Experimental results on Integer Programming and WASO-dis**

clined to select a connected group, where the optimal solution may be disconnected.

## 6. CONCLUSION AND FUTURE WORK

To the best of our knowledge, there is no real system or existing work in the literature that addresses the issues of automatic activity planning based on topic interest and social tightness. To fill this research gap and satisfy an important practical need, this paper formulated a new optimization problem called WASO to derive a set of attendees and maximize the willingness. We proved that WASO is NP-hard and devised two simple but effective randomized algorithms, namely *CBAS* and *CBAS-ND*, with an approximation ratio. The user study demonstrated that the social groups obtained through the proposed algorithm implemented in Facebook significantly outperforms the manually configured solutions by users. This research result thus holds much promise to be profitably adopted in social networking websites as a value-added service.

The user study resulted in practical directions to enrich WASO for future research. Some users suggested that we integrate the proposed willingness optimization system with automatic available time extraction to filter unavailable users, such as by integrating the proposed system with Google Calendar. Since candidate attendees are associated with multiple attributes in Facebook, e.g., location and gender, these attributes can be specified as input parameters to further filter out unsuitable candidate attendees. Last but not the least, some users pointed out that our work could be extended to allow users to specify some attendees that must be included in a certain group activity.

## Acknowledgment

This work is supported in part by US NSF through grants CNS-1115234, DBI-0960443, and OISE-1129076, US Department of Army through grant W911NF-12-1-0066, and National Science Council (NSC 101-2628-E-001-003-MY3), Taipei, Taiwan.

## 7. REFERENCES

- [1] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hofer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20:172–188, 2008.
- [2] V. Chaoji, S. Ranu, R. Rastogi, and R. Bhatt. Recommendations to boost content spread in social networks. In *WWW*, pages 529–538, 2012.
- [3] C. H. Chen, E. Yucesan, L. Dai, and H. C. Chen. Efficient computation of optimal budget allocation for discrete event simulation experiment. *IIE Transactions*, 42(1):60–70, 2010.

- [4] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. In *SIAM*, 51(4):661–703, 2009.
- [5] A. Costa, J. Owen, and D. P. Kroese. Convergence properties of the cross-entropy method for discrete optimization. *Operations Research Letters*, 35(5):573–580, 2007.
- [6] L. Dai, C. H. Chen, and J. R. Birge. Large convergence properties of two-stage stochastic programming. *Journal of Optimization Theory and Applications*, 106(3):489–510, 2000.
- [7] M. Deutsch and H. B. Gerard. A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology*, 51(3):291–301, 1955.
- [8] U. Feige, D. Peleg, and G. Kortsarz. The dense k-subgraph problem. *Journal of Algorithmica*, 29(3):410–421, 2001.
- [9] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *CHI*, pages 211–220, 2009.
- [10] D. F. Gleich and C. Seshadhri. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In *KDD*, pages 597–605, 2012.
- [11] N. Gupta, L. Kot, S. Roy, G. Bender, J. Gehrke, and C. Koch. Entangled queries: enabling declarative data-driven coordination. In *SIGMOD*, pages 673–684, 2011.
- [12] M. F. Kaplan and C. E. Miller. Group decision making and normative versus informational influence: Effects of type of issue and assigned decision rule. *Journal of Personality and Social Psychology*, 53(2):306–313, 1987.
- [13] M. Kargar and A. An. Discovering top-k teams of experts with/without a leader in social networks. In *CIKM*, pages 985–994, 2011.
- [14] C. Li and M. Shan. Team formation for generalized tasks in expertise social networks. In *SocialCom*, pages 9–16, 2010.
- [15] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: Inferring user profiles in online social networks. In *WSDM*, pages 251–260, 2010.
- [16] M. Mitzenmacher and E. Upfal. Probability and computing: Randomized algorithms and probabilistic analysis. Cambridge University Press, 2005.
- [17] R. Y. Rubinfeld. Combinatorial optimization, cross-entropy, ants and rare events. In S. Uryasev and P. M. Pardalos, editors, *Stochastic Optimization: Algorithms and Applications*, pages 304–358. Kluwer Academic, 2001.
- [18] H. H. Shuai, D. N. Yang, P. S. Yu, and M. S. Chen. Willingness optimization for social group activity. In *CoRR (1305.1502)*, 2013.
- [19] M. Sozio and A. Gionis. The community-search problem and how to plan a successful cocktail party. In *KDD*, pages 939–948, 2010.
- [20] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *Eurosys*, pages 205–218, 2009.
- [21] D. N. Yang, W. C. Lee, N. H. Chia, M. Ye, and H. J. Hung. Bundle configuration for spread maximization in viral marketing via social networks. In *CIKM*, pages 2234–2238, 2012.
- [22] M. Ye, X. Liu, and W. C. Lee. Exploring social influence for recommendation - a probabilistic generative model approach. In *SIGIR*, pages 671–680, 2012.